Analyzing the attitude of Indian citizens towards COVID-19 vaccine — A text analytics study

Praveen SV[*], Rajesh Ittamalla, Gerard Deepak

Department of Management Studies, National Institute of Technology, Trichy, India

ARTICLE INFO

ABSTRACT

Background and aims: The government of India recently planned to start the process of the mass vaccination program to end the COVID-19 crises. However, the process of vaccination was not made mandatory, and there are a lot of aspects that arise skepticism in the minds of common people regarding COVID-19 vaccines. This study using machine learning techniques analyzes the major concerns Indian citizens voice out about COVID-19 vaccines in social media.
Methods: For this study, we have used social media posts as data. Using Python, we have scrapped the social media posts of Indian citizens discussing about the COVID- 19 vaccine. In Study 1, we performed a sentimental analysis to determine how the general perception of Indian citizens regarding the COVID-19 vaccine changes over different months of COVID-19 crises. In Study 2, we have performed topic modeling to understand the major issues that concern the general public regarding the COVID- 19 vaccine.
Results: Our results have indicated that 47% of social media posts discussing vaccines were in a neutral tone, and nearly 17% of the social media posts discussing the COVID-19 vaccine were in a negative tone. Fear of health and allergic reactions towards the vaccine are the two prominent issues that concern Indian citizens regarding the COVID-19 vaccine.
Conclusion: With the positive sentiments regarding vaccine is just over 35%, the Indian government needs to focus especially on addressing the fear of vaccines before implementing the process of mass vaccination.

1. Introduction

During the mid of March 2020, WHO has announced the COVID-19 crisis as a pandemic global health crisis [1]. We are at the beginning of 2021, yet the normalcy hasn't returned, and many parts of the world are still in lockdown. The scariest aspect of this pandemic is, with the threat of the second wave being imminent, some experts believe the worst is yet to come. With countries like the United States of America and Germany, which records more COVID-19 cases now, compared to four months earlier, it can be inferred that the ongoing crises may well go beyond 2020. The year 2021 brings new hope to the Indian citizens as the daily surge of COVID-19 cases in India was under 20,000 cases for most parts of the year.

However, vaccinating the general public is one of the safest ways to halt any pandemics outbreak, including the current COVID-19 crises. The growing human rights concerns, anti-vaccine movement, and skepticism towards the vaccine and its effects may result in the process of vaccination becoming a complicated task.

However, no studies were conducted to analyze the common citizen's perception of COVID-19 vaccines. Our study focuses on analyzing Indian citizens' perspectives on COVID-19 vaccines. Our research focuses on answering the following questions.

RQ 1: What is the attitude of Indian citizens towards the COVID-19 vaccine? How it differs across various months of COVID-19 crises?

RQ 2: What are the major concerns voiced out by Indian citizens towards the COVID- 19 vaccine?

Previous studies have identified that several factors could lead an individual to develop their perspective towards vaccines and their effects [2,3]. Previous research has also indicated that people are more prone to reject the vaccine for a newly detected disease [4]. With the Indian government planning to vaccinate the citizens from the beginning of the year 2021, it is important to understand

* Corresponding author. Tanjore Main road, National highway 67, Tiruchirappalli, 620015, Tamil Nadu, India.
  E-mail address: praveennitt04@gmail.com (S. Praveen).

the general public's mental attitude towards the vaccine. This study will help government officials and policymakers understand the issues that need to be addressed before initiating the process of mass vaccination.

2. Methods

2.1. Data collection and data pre-processing

We have used social media posts of Indian citizens to understand common citizens' mental attitudes towards the COVID-19 vaccine and the effect of the vaccination during pandemic times. By using Python library Twint, we have scrapped tweets that contain the words 'Covid vaccine.' For this study, we have chosen Twitter as a data source because, after the outbreak of COVID-19, more and more individuals worldwide began to use Twitter as a platform to share their concerns about the pandemic and receive updates about COVID-19 [5]. From the previous studies conducted in the past, we can infer that social media is the best source from which we can understand the general public's behavior during unprecedented times [6—9].

For our analysis, we have collected tweets that contain the words "Covid vaccine." Python library Twint has an option of geographical filtering, and by using the option, we have collected tweets that belong to India. For this study, we have only chosen tweets written in the English Language. After eliminating the tweets that belong to other languages, we have finally chosen 73,760 unique English tweets for the study. To counter the damages that may be caused due to the unbalanced sample, we choose an equal number of tweets for each month in our corpus. After that, we performed the data cleaning process for the data in our corpus. We performed the process of removing the stop words, removing the numbers, punctuations, and hyperlinks that are not needed for our data analysis. Stop words are the words in the corpus that do not have any meaning of their own and are therefore not needed for the analysis. Stop words are generally articles and prepositions that do not have any meaning of their own and do not provide any merit. After removing the stop words from the corpus, we performed the process of stemming and lemmatizing to the data in our analysis.

Stemming is the process of eliminating the terminations of terms to identify their root type. On the other hand, lemmatization is the process of grouping different types of words together so that they will be grouped, thus reducing the dimensionality.

2.2. Research methodology

2.2.1. Sentimental analysis

The objective of study 1 is to understand the sentiments of Indian citizens towards the COVID-19 vaccine. We performed the process of sentiment analysis to understand the attitude of Indian citizens towards the COVID-19 vaccines. Sentimental analysis is a process by which the sentiment score of the data is determined after the analysis.

Sentiment analysis is described as "the automatic method to extract and analyze the subjective judgments on different aspects of an item or entity" [10]. Sentimental analysis is a machine learning process that involves the use of NLP (Natural Language Processing) with the objective of identifying the emotions the authors of a particular text expressed through his words [11]. Sentimental analysis at the beginning was done on the document level [12], then on the sentence level, and [13] on the phrase level [14,15].

Sentiment analysis computationally distinguishes and classifies opinion reflected in a particular section of text by the author about the subject that the premise is built upon. The sentiment analysis's main objective is to determine the rate of the polarity and by which

the tone expressed by the author in a particular corpus can be defined as positive, negative, or neutral. For this study, we used the computer programming language Python for collecting the tweets. Specifically, we have used the library twint for the process of data analytics. Some of the early studies conducted on sentiment analysis of Twitter data are [16,17]. For this study, we have used the python library Textblob to process the textual data. Textblob, using NLP (Natural Language processing) and advanced machine learning principles, analyze every word in the documents presented in the corpus and define the overall sentiments being projected as positive, negative, or neutral [18].

2.2.2. Latent Dirichlet Allocation (LDA)

The main objective of the first study is to understand the attitude of Indian citizens towards COVID-19 vaccines. However, a sentimental analysis study won't indicate the major factors or issues that shape the attitude of Indian citizens towards the COVID-19 vaccines. To understand the major aspects that lead Indian citizens to have the attitude, we have conducted study 2. We performed LDA (Latent Dirichlet Allocation) topic modeling in study 2 to understand the major aspects Indian citizen voices out that shapes their attitude towards COVID-19 vaccine. Latent Dirichlet Allocation topic modeling was first introduced in 2003 by Blei, Ng, and Jordan. Topic modeling is a process that summarizes a vast archive of texts by discovering the topics and themes hidden within a set of corpora by using a group of algorithms [19]. Before the introduction of LDA, Probabilistic Latent Semantic Indexing was used in deriving topics. The basic theory behind Probabilistic Latent Semantic Indexing is that the algorithms model each word in a document as a sample from a mixture model, where the mixture components are the multinomial random variables, and that can be represented as topics. However, one of the main disadvantages of Probabilistic Latent Semantic Indexing is its algorithms does not provide a probabilistic model at the document level [20].

LDA follows the "Bag of words" assumption, which represents a document as a mixture of latent topics in which a topic is a multinomial distribution over words. The core aspect of Latent Dirichlet Allocation (LDA) topic modeling is that the algorithms assume that all the documents in a data frame or corpus exhibit a similar set of topics which each of the documents exhibits a various probabilistic mixture of those topics. It works under the assumption that a specific group of words is prone to get associated with specific topics. Using LDA, one can discover latent topics from extensive unstructured data. To facilitate a better understanding of the identified issues, we used the library LDAvis.

3. Results

3.1. Study 1: Sentimental analysis

We have chosen 18,440 tweets across different months of the year 2020 to observe the attitude of Indian citizens towards the vaccine and its effects. It can be inferred from Table 1 that neutral and positive statements dominated the data frame. Nearly 47% of the social media posts in the last four months of 2020 about vaccines and their effects were in a neutral tone. 35% of the social media posts about the vaccine were in a positive tone, and 16.6% were in a negative tone. A close analysis of the table shows a slight increase in a positive and neutral tone and a considerable decrease in the negative sentiments as the number of cases piles up. There was an approximate 1.4% increase in positive tone during December comparing to the previous four months, and there was a 5.2% decrease in negative sentiments during December compared to October. It can be inferred that there is a strong correlation between the increase in positive sentiments towards the COVID-19 vaccine

Table 1
Sentimental analysis.

Month	Total Tweets	Neutral	%	Positiv e	%	Negati ve	%	Total cases
September	18,440	8672	24.5	6396	24.4	3372	27.4	6.3 M
October	18,440	8660	24.4	6512	24.9	3268	26.5	8.1 M
Novem ber	18,440	9084	25.6	6444	24.6	2912	23.6	9.4 M
December	18,440	8936	25.2	6764	25.8	2740	22.2	10.2 M
	73,760	35,352		26,116		12,292		

Note: The total in the first column refers to the total number of tweets belongs to a particular month. The fourth, sixth, and eighth columns refer to the percentage of positive, negative, and neutral tweets compared to the total number of positive, negative, and neutral tweets of all months in the data frame. For example, the month of September records 8672 neutral tweets, that is 24.5% of the total number of neutral tweets (n = 35,352) recorded across all the months in the data frame.

and an increase in the number of COVID-19 cases. The results of the sentimental analysis were given in Table 1. The graphical representation of the results of the analysis was given in Fig. 1 and Fig. 2.

3.2. Study 2: Topic modeling

Though sentimental analysis gave an insight into the general public's attitude towards the COVID-19 vaccine and its effects, it hasn't helped us understand the major issues that shape the general public's attitude. To understand the general public's concerns regarding the vaccine and its effects, we further performed Latent Dirichlet Allocation topic modeling for the tweets about the vaccine that has negative sentiments. The results of the topic modeling were given in Table 2.

4. Discussions and conclusion

4.1. RQ 1: What is the attitude of Indian citizens towards the COVID-19 vaccine? How it differs across various months of COVID-19 crises?

Although most common people's opinions on social media regarding vaccines and their effects were in neutral sentiments, only 35% of sentiments are positive, and this should be the concern for the government and policymakers. Unless or until the

government can convince most of its population that the results and the effects of the vaccine will be positive, the objective of vaccination cannot be achieved. With the positive sentiments regarding vaccine is just over 35%, the Indian government needs to focus especially on addressing the fear of vaccines before implementing the process of mass vaccination.

4.2. RQ 2: What are the major concerns voiced out by Indian citizens towards the COVID-19 vaccine?

Our findings reveal an interesting aspect that, despite COVID-19 having affected nearly 11 million people in India, a considerable number of Indian citizens still feel that the whole pandemic is exaggerated. This mentality will lead the citizens to reject the vaccine. Apart from that, skepticism over the nationality of the vaccine, skepticism over the vaccine trials, skepticism over health after taking the vaccine, the fear of death that the vaccine may cause, allergic reactions towards the vaccine, distrust over pharma companies, doubts regarding data provided by the vaccine companies, prevalence of numerous vaccines and concerns over choosing the safest, and the rush in providing vaccine were the other concerns shared by Indian citizens. Our analysis has shown that though the Indian general public is voicing some genuine concerns regarding the COVID-19 vaccine, some superstitious unproven conspiracy theories such as COVID-19 being exaggerated
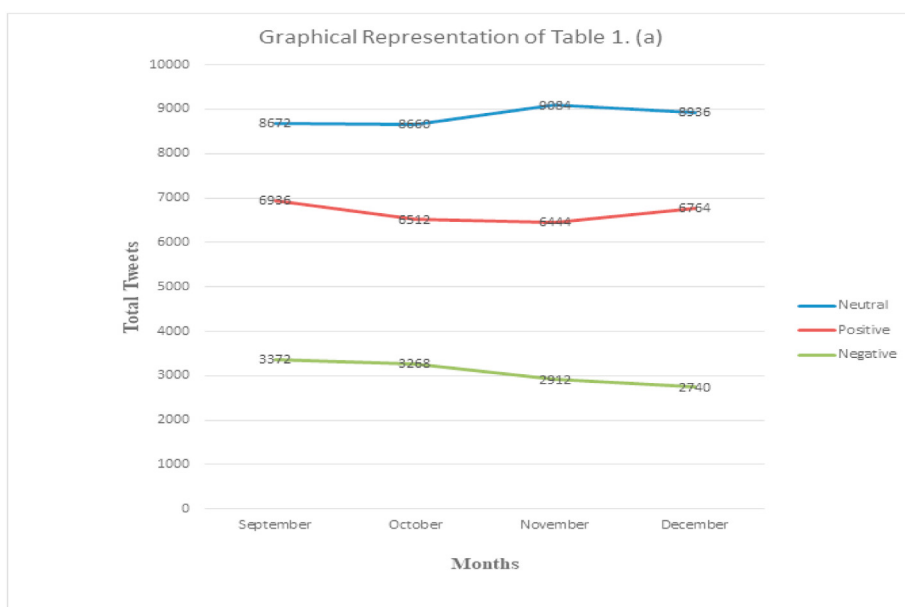

Fig. 1. Graphical representation of data for Table 1 (Months and Total number of tweets with positive, negative, and neutral sentiments for each month).
Note: Fig. 1 explains how the tweets with positive, negative, and neutral sentiments vary over the four months. For example, the green line represents the number of tweets with negative sentiments and how it differs for four months. As seen in the graph, in September, there is a total number of 3372 tweets with negative sentiments were recorded.
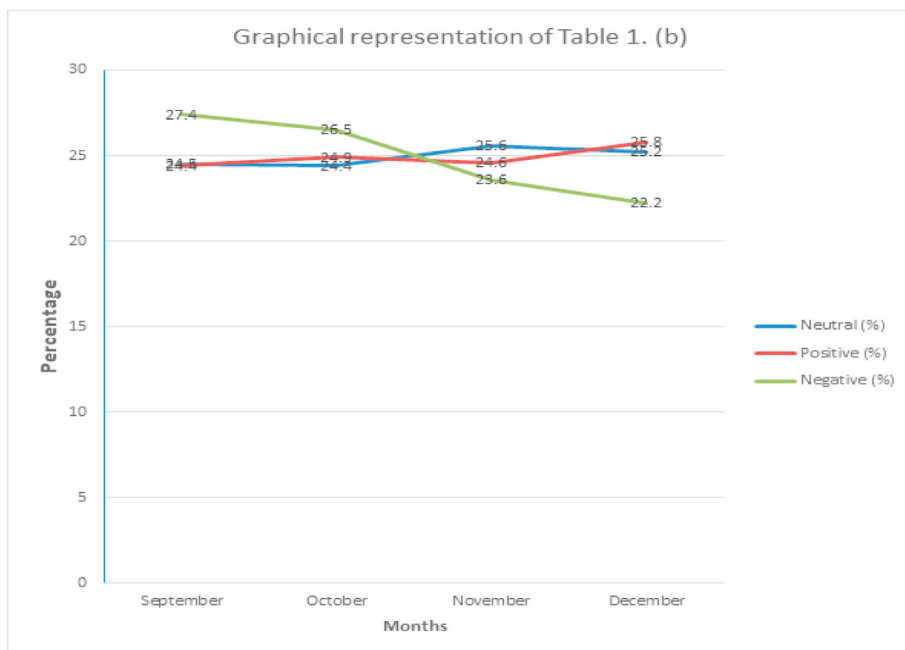
Fig. 2. Graphical representation of data for Table 1 (Months and percentage of positive, negative, and neutral tweets).
Note: Fig. 2 explains how the percentage of tweets with positive, negative, and neutral sentiments vary over the four months. For example, the green line in the graph represents the percentage of negative tweets for each month. As seen in the graph, the month of September alone records 27.4% of total tweets, with negative sentiments collected over all four months in the data frame.

Table 2
Latent dirichlet allocation topic modeling.

Topic	Top Words
Fear over health	take, health, vaccine, COVID-19, state, affect, jab
Allergic reactions	allergic, reaction, vaccine, dose, lab, China, patient
Fear of death	ready, news, COVID-19, vaccine, guess, heard, danger, dead
Skepticism over vaccine trails	get, say, trail, vaccine, test, actual, ill, won't
COVID-19 being exaggerated	COVID-19, vaccine, cure, laugh, exaggerate, freak, agenda
Skepticism over the nationality of the vaccine	efficient, Pfizer, Sputnik, trust, don't, government, Chinese, country
Negative feeling towards pharma companies	won't, don't, pharma, Bill, Gates, bad, CDC
Doubts regarding data	FDA, Pfizer, COVID-19, data, corrupt, safety, posit,
Various vaccines	pandemic, Fauci, AstraZeneca, Pfizer, Moderna, scientist, concern, safe
Rush in providing the vaccine	Will, year, need, school, rush, administration, effect, worry

and the dislike or disbelief toward certain vaccines because of the nationality were also echoed while discussing about the COVID-19 vaccine. Our studies show that a considerable percentage of the Indian population on social media does not trust government and pharma companies.

With recent research indicates that the immunity of even COVID-19 survivors is likely to stay in the system only for eight months [21], it is important for policymakers and the government to invest in educating the general public on the pros of vaccination and how important it is necessary to return to normal life. Considerable efforts should be taken by the governments, pharma companies, and NGOs in educating the general public about the vaccination program and the necessity of implementing it to return to normal life. Special emphasis should be given to ensure all the hoaxes and insecurities of the general public regarding the COVID-

19 vaccines are addressed so that it will motivate and encourage the general public to take vaccines.

References

[1] Who. WHO: coronavirus disease 2019 (COVID-19) situation report — 55. 2020. available at: www.who.int/docs/default- source/coronaviruse/situation-reports/20200315-sitrep55-COVID- 19.pdf?sfvrsn=33daa5cb_8.

[2] Zhu F-C, Li Y-H, Guan X-H, Hou L-H, Wang W-J, Li J-X, Wu S-P, Wang B-S, Wang Z, Wang L, Jia S-Y, Jiang H-D, Wang L, Jiang T, Hu Y, Gou J-B, Xu S-B, Xu J-J, Wang X-W, Chen W. Safety, tolerability, and immunogenicity of a recombinant adenovirus type-5 vectored COVID-19 vaccine: a dose-escalation, open-label, nonrandomised, first-in-human trial. Lancet 2020;395(10240):1845—54. https://doi.org/10.1016/s0140- 6736(20)31208-3.

[3] Wong CKH, Man KKC, Ip P, Kwan M, McGhee SM. Mothers' preferences and willingness to pay for human papillomavirus vaccination for their daughters: a discrete choice experiment in Hong Kong. Value Health 2018;21(5):622—9. https://doi.org/10.1016/j.jval.2017.10.012.

[4] Chou W-YS, Budenz A. Considering emotion in COVID-19 vaccine communication: addressing vaccine hesitancy and fostering vaccine confidence. Health Commun 2020;35(14):1718—22. https://doi.org/10.1080/10410236.2020.1838096.

[5] Wong Q. "Twitter's user growth soars amid coronavirus, but uncertainty remains. CNET 2020. www.cnet.com/news/twitters-user-growth-soars-amid-coronavirus- but-uncertainty-remains/.

[6] Wang B, Zhuang J. Crisis information distribution on twitter: a content analysis of tweets during hurricane sandy. Nat Hazards 2017;89(1):161—81.

[7] Chatfield A, Brajawidagda U. Twitter tsunami early warning network: a social network analysis of twitter information flows. 2012.

[8] Earle PS, Bowden DC, Guy M. Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys 2012;54(6).

[9] Buntain C, Golbeck J, Liu B, LaFree G. Evaluating public response to the Boston marathon bombing and other acts of terrorism through twitter. In: Tenth international AAAI conference on web and social media; 2016.

[10] Soleymani M, Garcia D, Jou B, Schuller B, Chang SF, Pantic M. A survey of multimodal sentiment analysis. Image Vis Comput 2017;65:3—14. https://doi.org/10.1016/j.imavis.2017.08.003. available at:.

[11] Li J, Hovy E. "Reflections on sentiment/opinion analysis", A practical guide to sentiment analysis. Cham: Springer; 2017. p. 41—59.

[12] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL; 2004.

[13] Hu M, Liu B. Mining and summarizing customer reviews. KDD; 2004.

[14] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase level sentiment analysis. ACL; 2005.

[15] Agarwal A, Biadsy F, Mckeown K. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams". In: Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009); 2009. p. 24—32.

[16] Bermingham A, Smeaton A. Classifying sentiment in microblogs: is brevity an advantage is brevity an advantage? ACM; 2010. p. 1833—6.

[17] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D, editors. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA); May 2010. p. 19—21.

[18] Tutorial: Quickstart. available at:, accessed, https://Textblob.Readthedocs.Io/En/Dev/Quickstart.Html. https://textblob.readthedocs.io/en/dev/quickstart.html. [Accessed 1 December 2019].

[19] Blei D, Lafferty J. Topic models. In: Srivastava AN, Sahami M, editors. Text mining: classification, clustering, and applications. Boca Raton, FL: CRC Press; 2009. p. 71—94.

[20] Blei D, Andrew N, Jordan M. Latent dirichlet allocation. J Mach Learn Res 2003;3:993—1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.

[21] Press Trust of India. Recovered COVID-19 patients last immunity for 8 months, raise hopes for vaccine: Study. India Today; 2020. https://www.indiatoday.in/coronavirusoutbreak/story/COVID-19-antibody-immunity- lasts-8-months-study-1752290-2020-12-23.